Reforming the UK Online Safety Bill to Protect Legal Free Expression and Anonymity

By Kir Nuthi and Mella Tesfazgi | August 29, 2022

In an attempt to respond to growing concerns about hate and awful activity online, the Online Safety Bill imposes duties of care—binding legal obligations for online services—that require online services to mitigate Internet harms such as hate speech, child predation, minor access to online pornography, self-harm encouragement, and more. While these policies are well intentioned and target credible harm and severe content online, the Online Safety Bill's loose definition of what constitutes "legal but harmful" content, overbroad scope, and general legislative overreach encroach on the civil liberties of all users—not just those in the United Kingdom. Specifically, the legislation undermines legal free expression, privacy, and anonymity. This report critically analyzes these challenges and provides alternative solutions that would minimize the Online Safety Bill's negative impact on important civil liberties and better balance its intended goals.

INTRODUCTION

The debate about how to moderate online speech has been at the forefront of the world's online policy discussions. While some policymakers and civil society groups believe that the Internet should remain a bastion of all legal speech—awful or not— or that others believe that the Internet should become a forum with a carefully calibrated balance between online safety, free expression, and privacy.[1] Several nations have aspired to legislate toward a balanced approach to legal free expression and online safety

through intermediary liability frameworks such as the United States' Section 230 of the Communications Decency Act and the European Union's e-Commerce Directive.[2] But for some proponents of greater online safety, these laws still fail to properly protect adults and children online and promote a digital environment ripe for misinformation.[3]

Through the proposed Online Safety Bill, the United Kingdom government has decided to take a different approach to content moderation to protect adults and children online. Originally pitched as a white paper in 2019 by the Department for Digital, Culture, Media, and Sport (DCMS) and then as a draft bill in 2021, the Online Safety Bill in its current form imposes a set of obligations—"duties of care"—on online services to monitor and remove various forms of both legal and illegal online content.[4] Services that fail to comply with the new rules face several sanctions, including fines of up to £18 million, or 10 percent of the services' qualifying worldwide revenue, whichever is higher.[5]

The modus operandi of the bill is to minimize legal but harmful content and mitigate illicit activity without infringing on civil liberties. But as policymakers have seen in other nations, it is impossible to protect both online safety and civil liberties without trade-offs.[6] By not acknowledging the implausible nature of a balance that eradicates legal and illegal online harm while preserving online freedoms, the Online Safety Bill falls into the same proverbial trap as its predecessors.[7]

While well intentioned, the bill's current approach to online safety is overly broad. The legislation does not clearly define what legal content online services should or should not moderate and creates mechanisms that severely undermine privacy and anonymity on the Internet.

The United Kingdom government can, however, amend the Online Safety Bill to address these shortcomings. To better protect legal free expression, it should revise the Online Safety Bill to take one of three approaches:

- Amend the bill only to restrict illegal content online and move certain harmful content from the lawful to unlawful category.
- Clearly define specific types of legal and illegal content it requires services to moderate to best position service providers to protect civil liberties such as freedom of speech.
- Codify intermediary liability protections for online services to proactively moderate content.

To better balance anonymity and safety online, it should revise the Online Safety Bill to:

- Remove age assurance or verification recommendations from the

proposal and prevent the Office of Communications (Ofcom) from prescribing this technology in the future.

- Protect encrypted communications from the scope of the Online Safety Bill.

THE ONLINE SAFETY BILL'S DUTIES OF CARE APPROACH

The Online Safety Bill is a sector-specific regulatory regime that specifically targets "user-to-user" and "search" online services that have the United Kingdom as either a target market or significant location of users.[8] A user-to-user service allows users to share user-created content that may be seen by other users, whereas a search services is a search engine or an online service that includes a search engine.[9]

The bill places a legal responsibility on services to moderate illegal and certain types of legal but harmful content on their platforms. Harm, as defined by the Online Safety Bill, refers to content that causes physical or psychological harm based on the nature, dissemination, and distribution of user-generated content.[10] It occurs whenever, as a result of the content, individuals act in a way that results in harm to themselves or others, or increases the likelihood of harm to themselves or others.[11]

User-to-user services regulated under the bill are expansive, as the bill only exempts the following: User-to-user services are exempt if their only user-created content is emails, SMS messages, MMS messages, aural communication, reviews, or some combination of these types of content, so long as these services do not contain pornographic content and have links to the United Kingdom.[12] Links with the United Kingdom are defined as a "significant number of United Kingdom users" or the United Kingdom being a "target market" for the online service.[13] All other user-to-user services are covered under the bill and not only include traditional social media and forums but also over-the-top messaging platforms such as iMessage, Signal, and WhatsApp—messaging platforms that either don't need phone numbers to work or that use phone numbers for two-factor authentication but otherwise operate and send user-to-user content over the Internet.[14]

TYPES OF CONTENT THE ONLINE SAFETY BILL OBLIGATES ONLINE SERVICES TO MODERATE

The Online Safety Bill creates these legal obligations through its duties of care placed on online services to address illegal content, content harmful to children, and content harmful to adults.

Illegal Content

Services have a duty to carry out risk assessments to categorize the risk users may face of encountering illegal content on their platform.[15] The bill

defines "illegal content" as "content that amounts to a relevant offense."[16] The bill also specifies "priority illegal content" as a component of illegal content, which includes child sexual exploitation and abuse (CSEA), terrorist content, and other priority offenses.[17] Priority offenses include assisting suicide, threats to kill, intentional harassment that causes fear of violence, racially or religiously aggravated public order offenses, offenses regarding firearms, sexual offenses, assisting unlawful immigration, proceeds of crime or fraud, and more.[18] Illegal content includes priority illegal content but also an offense under any United Kingdom law wherein "the victim or intended victim is an individual or (individuals)."[19]

Online services also have a duty to effectively protect all users from encountering, generating, and sharing illegal content.[20] Online services must do this moderation throughout all areas of their service through proactive technology, content moderation, and other policies on their platforms.[21] Unfortunately, due to the scope of services covered and the requirement of proactive technology, duties such as this within the Online Safety Bill will come at the expense of encrypted communications.

Content Harmful to Children

The Online Safety Bill compels online services to use proactive technology and risk assessments to prevent child access to these three categories of legal but harmful content.[22]

Under the Online Safety Bill, content is described as harmful to children when it is primary priority content, priority content, or content that "presents a material risk of significant harm to an appreciable number of children in the United Kingdom."[23] After the fact or in secondary legislation, the secretary of state for DCMS will designate content into the primary priority content that is harmful to children if it's of material risk and appropriate for the child safety duties, while priority content that is harmful to children is specified by the secretary of state for DCMS if "there is a material risk of significant harm to an appreciable number of children presented by content of that description that is regulated user-generated content or search content."[24] This makes the three categories of content harmful to children nearly indistinguishable until clarified by the secretary of state for DCMS. The secretary of state for DCMS cannot designate content into these categories if the risk of harm is derived from the content's potential financial impact, safety or quality of goods featured, or how services are performed.[25]

Within the duty to operate these proactive and proportionate measures, the Online Safety Bill is careful to not prescribe but instead recommend age verification and other age assurance measures to prevent child access to harmful or illegal content. Age assurance refers collectively to age verification and age estimation approaches used to prevent children from

accessing "adult, harmful, or otherwise inappropriate content" online.[26] Unfortunately, the prescription of age verification and age assurance measures is not without additional user privacy and safety concerns as discussed later in the paper.

Content Harmful to Adults

Similar to the safety duties to protect children online, these duties of care cover content that is "legal but harmful" to adults—priority content that will be specified later by the secretary of state for DCMS and content that presents a "material risk of significant harm to an appreciable number of adults in the United Kingdom."[27] Unlike the safety duties to protect children online, these apply only to high-risk user-to-user services—defined in the bill as Category 1 services (and clarified below).

These duties of care within the Online Safety Bill compel online services to use proactive technology and risk assessments to prevent adult access to any legal but harmful content covered by their terms and conditions.[28] Unlike the duties to protect children online, the duties to protect adults online require services to create clear terms and conditions that specify how they deal with each type of priority content or other content harmful to adults, how they conduct risk assessments for this content, and how they moderate this content.[29] If an online service's terms and conditions do not address types of legal but harmful content or content that is not designated but still harmful to adults, these services must notify Ofcom.[30] It is unclear whether services must mitigate the risk of this content or how the definition of and duties governing content harmful to adults will change under the discretion of Ofcom and the secretary of state for DCMS.

For legal but harmful content for adults, services must specify how they will apply their terms and conditions to either take down, restrict access to, limit the recommendation or promotion of, or recommend and promote these types of content.[31] Three of these options—taking down, restricting access, and limiting the recommendation of legal but harmful content—would minimize and prevent the content from reaching users while one—choosing to recommend and promote the content—offers services the ability to keep the content up if they explicitly explain so in their moderation practices.[32]

HOW THE ONLINE SAFETY BILL CATEGORIZES ONLINE SERVICES

The Online Safety Bill also splits types of companies covered by its duties of care into two categories, of which the thresholds for categorization are up to the secretary of state for DCMS and are based on number of users, functionality, and other relevant indicators.

Category 1 (High-Risk User-to-User Services)

This encompasses user-to-user services of high risk for the content moderated under the Online Safety Bill. Category 1 will likely refer to user-to-user services with the largest audiences and with a range of high-risk features. Category 1 services will face safety duties to protect adults, risk assessments regarding adult safety on their services, as well as duties to protect content of democratic importance, and journalistic content.[33] Category 1 services also face a duty regarding the potential prevalence of fraudulent advertising on the platform.[34]

Category 2 (Lower-Risk User-To-User Services and Search Engines)

This section is split into two subcategories: Category 2A encompasses regulated search services and a combination of user-to-user and search services considered to be of lower risk. Category 2B encompasses regulated user-to-user services considered to be of lower risk, as defined by the Ofcom registry based on threshold conditions the secretary of state for DCMS will specify later through secondary clarification.[35] Category 2A services also face a duty regarding the potential prevalence of fraudulent advertising on the platform.[36]

Table 1: Types of Content and Services Regulated by the Online Safety Bill

	Category 1	Category 2A	Category 2B
Definition	Higher-Risk User-to-User Services	Search Engines	Lower-Risk User-to-User Services
Duties for Illegal Content	Covered	Covered	Covered
Duties for Content That is Legal But Harmful (Children)	Covered	Covered	Covered
Duties for Content That is Legal But Harmful (Adults)	Covered		
Duties to Protect Content of Democratic Importance	Covered		
Duties to Protect Journalistic Content	Covered		
Duties to Prevent Fraudulent Advertising	Covered	Covered	

THE ONLINE SAFETY BILL'S ENFORCEMENT MODEL

OFCOM IS POSITIONED AS THE KEY ENFORCER OF THE ONLINE SAFETY BILL

Ofcom—the United Kingdom's broadcast, telecom, and postal regulatory agency—and the secretary of state for DCMS in consultation with one another will enforce the duties of care delineated within the Online Safety Bill.[37]

Ofcom will have the power to create codes of practice—guidelines vetted by multiple subject matter experts—that will recommend measures for complying with the Online Safety Bill's duties of care, which will then be submitted to the secretary of state for DCMS to lay before Parliament for approval.[38] If approved, these codes of practice become the guidelines by which Ofcom can judge whether an online service provider is complying with the necessary duties of care.[39] The secretary of state for DCMS can direct Ofcom to modify codes of conduct for reasons of public policy, national security, or public safety.[40] Similarly, Ofcom can also propose minor amendments without laying them before parliament if the secretary of state for DCMS agrees that they are minor and the consultation is unnecessary.[41]

Ofcom will also have the power to monitor compliance, issue notices and decisions that potentially compel online services to use proactive technology, investigate policy breaches, charge senior managers criminally, and issue fines to online services.[42]

In fact, Ofcom can impose significant financial and criminal sanctions on relevant services found in breach of the new regime. Penalties can amount to up to 10 percent of qualifying worldwide revenue or £18 million, whichever is highest.[43] And since the bill requires services to name senior managers charged with ensuring compliance, managers who fail to comply with Ofcom's audits or woefully mislead the regulator can individually face criminal charges.[44]

RISK ASSESSMENTS BY OFCOM AND ONLINE SERVICES

One of the Online Safety Bill's duties of care outside content moderation requires online services to carry out risk assessments that determine whether content hosted on their platform contains words or images that violate the new policy.[45] These risk assessments will be conducted against Ofcom's codes of conduct to encourage the use of practices and minimize content recommended or mentioned in these codes of conduct. Similarly, these risk assessments will be further clarified by Ofcom guidance and Ofcom's own risk assessments of each service in the sector. Ofcom will conduct its own analysis if services comply with the duties of care and will look at the risk assessments done by the online services. Ofcom's risk

assessments will focus on content harmful to adults and children, illegal content, and "non-designated content that is harmful to adults."[46]

NEGATIVE IMPACTS ON LEGAL FREE EXPRESSION

As seen through the definitions of harmful content and the enforcement mechanism, the Online Safety Bill creates standards that are subject to redefinition. With so many details left to later legislation, the Online Safety Bill is unclear on what online services should monitor and remove. At best, this will confuse companies regarding what content they should be proactive about and will cause them to struggle to consistently apply such a subjective standard. But at worst, this potential constant redefinition will push companies to over-moderate for fear of hefty fines and criminal charges for their staff.

VAGUE DEFINITIONS OF "LEGAL BUT HARMFUL" CONTENT LEAD TO BAD MODERATION

One major problem with the Online Safety Bill is its treatment of legal but harmful content. The term "legal but harmful" has become a key part of the debate about the Online Safety Bill even though this phrase does not actually appear in the legislative text. Instead, the term refers to two types of content covered in the Online Safety Bill: content deemed harmful to children and content deemed harmful to adults, the latter of which only applies to Category 1 services. The Online Safety Bill obligates services to moderate this type of content to protect Internet users in the United Kingdom.

According to the DCMS response to the Joint Bill Committee draft report for the Online Safety Bill, content that is legal but harmful to adults may encompass "racist and misogynistic abuse which doesn't meet the criminal threshold" while content that is legal but harmful to children may encompass "grooming, bullying, pornography and the promotion of self-harm and eating disorders."[47] The United Kingdom government is careful not to create a priority list in the legislation that delineates content that could fall under this category. Instead, content falling under the scope of legal but harmful content that online services must proactively monitor and remove will be determined in later legislation and will be subject to parliamentary approval.[48]

This lack of clarity is further compounded by how the secretary of state for DCMS and Ofcom may continuously adjust and redefine what legal but harmful content to which the Online Safety Bill's duties of care apply.

Potential Causes for Redefinition of Legal but Harmful Content for Children

The secretary of state for DCMS has the ability to consult with Ofcom to

"specify a description of content" for any regulated user-generated content or search content that can be considered of "material risk of significant harm to an appreciable number of children."[49]

"Material risk of significant harm to an appreciable number" is a vague phrase the legislative text does not further clarify. The terms "material," "significant," and "appreciable" are subjective and up for interpretation. "Material" illustrates the importance of the risk in the decision calculus of the secretary of state for DCMS, while "significant" illustrates the importance and noticeability of harm in the decision calculus and "appreciable" illustrates the importance and noticeability of the fraction of the population affected.[50] The use of three subjective words enables the secretary of state for DCMS in consultation with Ofcom to be equally as subjective when justifying content moderation decisions. Justifying content moderation decisions without including any qualitative or quantitative metric is not a replicable standard. Instead, the ability of the secretary of state for DCMS in consultation with Ofcom to redefine legal but harmful content for children is open to an inherently subjective interpretation.

Potential Causes for Redefinition of Legal but Harmful Content for Adults

The secretary of state for DCMS has the ability to consult with Ofcom to "specify a description of content" for any regulated user-generated content or search content that can be considered of "material risk of significant harm to an appreciable number of adults."[51]

Once again, the definition of what can or can't be redefined by the secretary of state for DCMS in consultation with Ofcom is open to an inherently subjective interpretation, adding a second subjective definition judged by the secretary of state for DCMS into the mix. But unlike how the safety duties regarding children can affect all regulated online services, this subjective definition will only apply to Category 1 services and must be dealt with consistently as specified in these services' terms and conditions.

These two subjective definitions that comprise legal but harmful will only muddle the enforceability and viability of the Online Safety Bill. Without definitional clarity, online services will face a national regulation defining content moderation that may result in politicized content moderation or over-moderation.

Content Moderation Could Become Politicized

Because of how the secretary of state for DCMS can in conjunction with Ofcom adjust and redefine the specific content on which the Online Safety Bill's duties of care hinge, the standards of what can and can't be allowed online lie in the hands of a political appointee.[52] By placing this much decision-making power in the secretary of state for DCMS, the content that

companies proactively monitor and remove and the codes of practice with which these companies must comply can and will change depending on the party in government.

The prime minister decides on secretary of state appointments, such as the one for DCMS, from a pool of members of Parliament in the House of Commons and members of the House of Lords.[53] Power shifts between political parties and political coalitions, changes in party leadership, prime minister resignations, resignations from secretaries of state, and outright dismissals can change who has decision-making power over the Online Safety Bill and what constitutes legal but harmful. So, while the Online Safety Bill was introduced under Conservative Party leadership and the premiership of Boris Johnson, it's likely to be implemented by different Conservative Party leadership and taken advantage of by any government afterwards. If the political tide is in the Conservatives' favor, then the definition of legal but harmful can be adjusted and molded into the political needs and subjective definition of the Conservative Party. If the tide sways to the Labour Party, then Labour ministers can take advantage of its powers when the next Labour government takes office. A Labour member of Parliament has already come out implying that extremist content from "incels" and "climate deniers" should be included in the definition of legal but harmful content.[54] The fact that Labour and Conservative members of Parliament are already discussing extensions of what is "priority content that is harmful to adults" highlights that there will be never-ending redefinition of legal but harmful content.[55]

United Kingdom users could face stark changes between political appointments and leadership changes. Online discourse on contentious issues that challenge the sitting party, such as the Brexit referendum in 2016, can be decided as harmful depending on the whims and needs of appointed political leaders. In practice, this constant redefinition could change what the Internet in the United Kingdom looks like on a political timeline.

Furthermore, this redefinition can and will have extraterritorial effects. Online services seeking to minimize liability under the Online Safety Bill and maintain efficiency may integrate product changes that reflect United Kingdom law for their global user-base, subjecting non-United Kingdom Internet users to similar moderation constraints. Thus, legislation that lacks definitional clarity and is susceptible to re-interpretation is ill-advised, as these faults found in the Online Safety Bill could spread internationally.

Content Users and Services That Prefer to Remain Online Could Be Over-Moderated

Further, the changing definition of what content is and isn't allowed online is likely to compel online services to implement the strictest content

moderation practices that comply with Online Safety Bill regulations. Regardless of potential redefinition, the current definition of legal but harmful is vague and will likely cause online platforms to over-moderate for fear of penalties, such as criminal charges for their staff and fines for noncompliance. That means services could over-moderate lawful speech that they, and their users, would prefer to allow to remain online.

Vague definitions will result in over-moderation because, when it comes to legislation, more context—not less—is necessary. If Ofcom has "reasonable grounds to believe" an online service is failing to comply with one of the duties of care, it can give provisional notices of contravention and start the enforcement process.[56] This reasonable belief standard is a subjective and conditional definition that is ripe for misinterpretation. Given how content moderation on the Internet works and the scale of content posted, any "reasonable grounds to believe" standard will be difficult to apply consistently. In 30 minutes alone, Facebook takes down more than 600,000 pieces of content, YouTube over 250,000, and TikTok 18,000.[57] Given the social media renaissance over the last two decades that has attracted a user base accounting for over 58 percent of the world's population, these numbers are likely to remain large, or grow to be even more substantial.[58]

Online services currently struggle to differentiate between various types of content, such as a user encouraging self-harm versus someone raising awareness around postpartum depression, photos affirming breastfeeding versus exploitative nudity, and political free speech versus misinformation.[59] Penalizing online services for making the wrong call will make platforms err on the side of over-moderation. In addition, online services will be forced to choose between focusing on the context behind user-created content, the Online Safety Bill's duties of care as they stand, and the potential for situations to occur that are unprecedented and don't match the content regulated in the bill or predicted in the services' terms and conditions. At best, these services will be able to accurately guess what content to moderate in compliance with the Online Safety Bill. But at worst, these services could impede swaths of legal free speech wrongfully deemed harmful as they navigate and attempt to enforce vaguely defined rules. And if companies have never seen a potentially harmful piece of content before, it could mean discussions on important topics have to be taken down before online services, users, and policymakers even get a chance to discuss what's best for the public.

Couple this with how online services are required to carry out risk assessments over whether they've successfully prevented content that violates the Online Safety Bill and the end result is clear: Online services will focus on avoiding providing assessments to Ofcom that show they do not "effectively mitigate and manage the risks of harm to individuals," as

identified in the most recent risk assessments.[60] Instead, services will effectively over-moderate to avoid enforcement penalties for noncompliance.

JOURNALISTIC VALUE AND DEMOCRATIC IMPORTANCE EXEMPTIONS FACE MODERATION CONCERNS

The Online Safety Bill makes a distinction between contentious speech that should be removed and important but potentially contentious speech that provides benefits to the public. For this reason, the Online Safety Bill creates provisions to preserve speech that by its duties of care may be considered priority harmful but is of democratic importance or journalistic value. Unfortunately, even these exemptions face the same over-moderation concerns, as their lack of clarity makes it likely that online services will struggle with subjective and contextual nuances that are not encompassed or clarified by the bill's text.

Sections 15 and 16 of the Online Safety Bill define these "duties to protect journalistic content" and "duties to protect content of democratic importance" as only applicable to Category 1 services, the higher-risk user-to-user-services.[61]

Duties to Protect Journalistic Content

These are designed to ensure journalistic free expression for United Kingdom-linked content that is either news-publisher or user-created content generated for the purposes of journalism.[62] Links with the United Kingdom are defined as a "significant number of United Kingdom users" or if the United Kingdom is a "target market" for the online service.[63]

The journalistic value of content must be taken into account and protected when making content moderation decisions.

Duties to Protect Content of Democratic Importance

This provision exempts United Kingdom-linked content on online services from the Online Safety Bill's other duties of care if it's from a recognized news publisher or is regulated user-created content that is specifically contributing to the "democratic political debate in the United Kingdom or a part or area of the United Kingdom."[64]

User-Created Content with the Purpose of Journalism and Democratic Debate

The question of potential over-moderation and definitional clarity centers around the second type of content described in these exemptions: user-created content generated for the purposes of journalism and democratic debate. Colloquially, this type of content is citizen journalism—the collection, analysis, opinions, editorials, and dissemination of real-time news and information by public citizens.[65]

Unlike news and opinions disseminated by a "recognised news publisher," content disseminated by citizen journalists does not have the formalized processes, staff assistance, policy procedures, and standards codes news outlets have—which are characteristics the Online Safety Bill uses to categorize outlets as qualifying for the journalistic value exemption.[66] In fact, this is the distinction that drives potential fears of over-moderation of citizen journalism.

Distinguishing between a citizen journalist and a regular Internet user is difficult, and the Online Safety Bill fails to recognize this or clarify what content falls under legal but harmful and what content falls under "journalistic value" or "democratic importance." This lack of distinction leaves online services without a clear guideline for distinguishing between user-created content that must be taken down and content that is exempt from this regulation. The Online Safety Bill's explanatory notes do include examples of "democratic importance," but these examples are few and only capture a subset of content.[67] Examples of promoting or opposing policies and parties for "democratic importance" only provide some examples that online services know to protect when moderating and fail to provide a holistic test these services can use to ensure compliance with the Online Safety Bill.[68] These same explanatory notes fail to even provide similar examples for "journalistic value," limiting it to seemingly any content "generated by news publishers, freelance journalists and citizen journalists."[69]

Since the bill does not provide this clarity over what is and isn't citizen journalism, platforms can easily fall into one of two traps: over-moderation or under-moderation.

Services Could Over-Moderate Content with Journalistic or Democratic Importance

Without clarity over what regulated user-created content is and isn't considered of journalistic and democratic importance, online services can easily miscategorize user-created content as lacking the qualities necessary for these exemptions before applying their content moderation practices.

Flexible and agile content moderation practices ensure that user-created content is not censored. Social media already struggles to moderate social issues and movements in real time.[70] Facing an ever-changing world with strict yet vague restrictions could suppress content United Kingdom users might find poignant or personally important.

For instance, social media was critical in the discussions surrounding the shooting of Mark Duggan.[71] If decisions by social media were not made quickly to keep up these posts, debates about the police's response and

use of force could have been unfairly censored and taken down. It was eyewitness accounts that told the full story, not the Independent Police Complaints Commission, which admitted that it might have been inadvertently providing misleading information to journalists.[72] Similarly, the discussions of women's safety after the horrifying murder of Sarah Everard also used social media as a tool for social change.[73] Women from across the United Kingdom and around the world discussed their worries about public safety, organized a vigil to "Reclaim These Streets," and organized "Kill the Bill" protests to fight a proposed law that would give more powers to the police.[74] If social media were forced to deal with the duties of care over what is legal but harmful, these services would have to quickly categorize the user-created content before applying content moderation practices. Categorizing this content too quickly could lead to content of democratic and journalistic importance being over-moderated, thereby removing the possibility of conversations such as these in the future.

In fact, the struggle to navigate duties of care against legal but harmful content and exemptions for "journalistic value" and "democratic importance" won't just affect United Kingdom Internet users. Yes, these provisions will likely lead to over-moderation of the United Kingdom Internet space and delay United Kingdom Internet users' ability to receive and share citizen journalism. But these provisions could easily snowball if online services choose to apply Online Safety Bill rules extraterritorially in order to minimize their compliance burden or prevent potential debates over whether international content is United Kingdom-linked. This could consequently limit the ability of citizen journalists to affect change globally.

Given the global nature of journalism and democratic debate, this will affect journalism internationally and have extraterritorial effects. News stories about other countries are often United Kingdom-linked and, as they can be read by a "significant number of United Kingdom users" and many news outlets and users treat the United Kingdom as a "target market" for the online services.[75]

A citizen of a war-torn country may post distressing images intended to raise global awareness of the situation at home. Overt violence, blood, or gore within the post could trip the definition of legal but harmful content. But if services fail to acknowledge the democratic importance behind this seemingly harmful post, they will likely remove speech vital to sustaining a healthy democracy. Or take recent social movements that have advocated for needed social change globally. Movements such as Arab Spring, #MeToo, BlackLivesMatter, and the Ukrainian support effort started as users of online services working together to gain their footing online and create simultaneous online and offline conversations and solutions.[76] These are examples that need flexible and agile content moderation

practices that will not censor the user-created content. Globally, these are considered movements of journalistic and democratic importance now, but if created under the regime of the Online Safety Bill, could have faced significant hurdles—more than what they currently face. Social media already struggles to moderate social issues and movements in real time.[77] Facing an ever-changing world with strict yet vague restrictions could suppress the next social movements.

Services Could Under-Moderate Content Under the Guise of Exemptions

Online services could also choose to use the vague and overbroad definitions of "journalistic value" and "democratic importance" to potentially under-moderate content. These definitions leave a significant amount to interpretation and could lead to many users choosing to identify as citizen journalists for the purpose of bypassing content moderation regulations. Not all who claim to be journalists and pundits demonstrate journalistic integrity, and social media reduces the barriers to entry for anyone to post their political opinions online. This phenomenon occurred in the rise of QAnon in the United States—a far-right, anti-Semitic political conspiracy movement that gained traction on social media.[78] QAnon garnered a large online following, with its ideas perpetuated by online pundits and even news publishers such as the *Epoch Times*.[79] Prominent online services took aggressive steps to stem the flow of QAnon content after the January 6 assault on the United States Capitol.[80]

QAnon is an extreme example, but it highlights a flaw in the moderation scheme for the Online Safety Bill's exemptions. In order to be excluded from the exemptions and henceforth regulated, outlets that perpetuated QAnon conspiracies would have to be proscribed as terrorist organizations and all commentary would become regulated content.[81] This designation could lag behind based on the fact that these "excluded entities" would have to be added to the list in the Terrorism Act 2000 and would require legislative action.[82] Similarly, in order to regulate against citizen journalists who perpetuated QAnon conspiracies, their content would have to either be removed from the "democratic importance" or "journalistic value" exemptions or be regulated under the duties of care regarding illegal content. This will lag behind based on the need for either Ofcom and the secretary of state for DMCS to amend the list of priority illegal content or online services to feel comfortable taking significant decision-making power under an Online Safety Bill regime.

NEGATIVE IMPACTS ON PRIVACY AND ANONYMITY

The Online Safety Bill creates a government-compelled monitoring standard for online services and targets a large swath of online services with its duties of care. As a result, the bill's attempt to reduce online harms

will simultaneously result in government-sanctioned surveillance on many online services, including encrypted communications. The recommendation of age assurance measures within the bill coupled with the bill's directive for online services to use proactive technology and the scope of services covered will erode United Kingdom residents' privacy and make them more vulnerable to bad actors online.

RECOMMENDING AGE ASSURANCE CREATES PRIVACY AND SECURITY VULNERABILITIES

Age assurance measures, as encompassed within the Online Safety Bill, are found in the duties to address content harmful to children. The bill is also careful to not mandate age verification but instead recommend "age verification, or another means of age assurance" as a way for online services to protect children of all age groups from harmful content or "primary priority content that is harmful to children."[83]

Age assurance broadly refers to processes, software, and other means online services can use to ensure that users are of a certain age.[84] The United Kingdom Information Commissioner's Office issued an opinion on how to comply with the Age appropriate design code—the United Kingdom's Children's code to protect children's data under current data protection law such as the United Kingdom General Data Protection Regulation—in which it clarified the working definition of age assurance.[85] This opinion clarified age assurance as the following:

> *"Age assurance" refers collectively to approaches used to:*
>
> - *Provide assurance that children are unable to access adult, harmful or otherwise inappropriate content when using [information society services]; and*
> - *Estimate or establish the age of a user so that [information society services] can be tailored to their needs and protections appropriate to their age.*[86]

The Information Commissioner goes further on to describe age assurance as split into two options: age verification and age estimation.[87] Age verification requires significant proof that a user is not underage, using what are considered "trusted, verifiable records of data."[88] Age estimation can use a variety of different approaches but often uses algorithmic means to categorize age demographics.[89]

The Online Safety bill allows Ofcom and the secretary of state for DCMS to mandate online services put any content arbitrarily deemed as harmful to an "appreciable number of children in the United Kingdom" behind an age assurance wall.[90] Similarly, the Online Safety Bill compels online services to use age verification to ensure children cannot encounter pornographic

content.[91]

Age assurance measures could range from an easily circumventable date of birth form to more complicated measures such as using users' passport numbers, driver's license scans, or other forms of government identification as entry to content and services the government deems unsuitable for children's eyes. Many online services hosting age-restricted content already employ age gates, but since some are easier to bypass than others, all services may be compelled to employ the stricter age verification method to meet this obligation, regardless of a service's risk profile. As a result, disclosing personally identifiable information to access online content, such as legal online pornography that could be previously accessed anonymously or semi-anonymously, could become the norm.

In this sense, the age assurance measures recommended within the Online Safety Bill can negatively affect users who are risk averse over where they put their personally identifiable information or who must remain anonymous for their own well-being. For children, many do not have government-issued IDs so it would be impossible to use age verification to control their access to services—and many services would have to turn to age estimation tools.

Age Verification Can Expose Users' Personal Information to Security Risks

Tying personally identifiable information in these ways could endanger user privacy as services obtain and try to protect new kinds of sensitive user information. Although some platforms have proposed less invasive methods for verifying user age, others are exploring more invasive options such as analyzing subjective language or even requesting biometric data.[92] In response to growing concerns about the accessibility of harmful content, Google stated that it would ask users of its video-sharing platform YouTube to verify their age by uploading credit card information before they could watch adult-only content.[93] Meta has started developing programs to look for signs that a user is lying about their age, such as spotting when someone claiming to be over 21 receives a message about her Bat Mitzvah.[94] But scanning text communications may prove ineffective, as context and nuance can be lost in translation, and users can be often hyperbolic in speech.

Requiring users to disclose more personal information online can undermine user privacy and anonymity. Bad actors can use personal information—if not protected sufficiently—to exploit and extort individuals. In fact, data breaches of personally identifiable information are already too common in a world without heightened age verification and age assurance. To date, millions of users' personal information has been compromised in hacks of Equifax, Cambridge Analytica, and Target—sites that don't

necessarily have age verification requirements—just to name a small sampling of data breaches.[95] And similar breaches have occurred with Ashley Madison—an adult extramarital affairs site that did require users to be above a certain age—when a 2015 cyberattack compromised users' names, contact information, and other personally identifiable information.[96]

Age Verification Can Expose Marginalized Communities That Rely on Anonymity

Ashley Madison, from a moral standpoint, is a contentious example of users losing the right to anonymity. Some argue that those committing adultery should not expect a right to privacy if they solicit digitally, while some believe in a form of privacy absolutism that ensures everyone has a right to anonymity. And many remain somewhere in the middle. But what the Ashley Madison scandal—and many scandals like it, such as GamerGate—highlight is the risk to people's livelihoods and emotional well-being when breaches and leaks compromise their online anonymity. The company offered users a paid option that would wipe them from Ashley Madison's database and search results, but a leaked company document shows that credit card information and emails still remained in its database.[97] As a result, when the site suffered a hack, thousands of users who were misled into believing their information would be de-identified were stripped of their anonymity. Users were left feeling distraught and in some extreme cases resorted to taking their own lives.[98]

If this scenario is repeated again, but with marginalized communities such as dissidents, human rights activists, or abuse survivors, the situation becomes much more dire and morally catastrophic.[99] In fact, any tool that can potentially compromise a user's right to anonymity—such as age verification through government-identification or biometrics—disproportionately affects those who need online anonymity in order to feel safe. Forcing members of these communities to identify their age or other personally identifiable information through age assurance measures could pose significant risks to their safety both on and off the Internet.

A privacy breach revealing members of an anonymous LGBT+ forum could lead to job loss, strained personal or professional relationships, or even self-harm for members still choosing with whom to share their sexual orientation.[100] In fact, LGBT+ youth rely on anonymity to avoid potential family persecution, being outed without their consent, and to find resources they need to remain safe.[101] Activists rely on pseudonyms to escape online threats and could suffer violent consequences if bad actors got ahold of their personally identifiable information.[102] Dissidents rely on anonymity as well as end-to-end encryption to avoid persecution and government surveillance under authoritarian regimes.[103] A hack could even endanger the lives of abuse survivors using pseudonyms to share stories,

resources, and other information with their community. Without anonymity, domestic violence and abuse survivors would lose tools they use to protect their online lives from their abusers.[104]

These are just a small smattering of examples of why anonymity is critical online to marginalized communities. And this is exactly why age assurance—or any method that puts more verifiable personally identifiable information online—risks far-reaching consequences for users on the margins. Age assurance in all these cases could force marginalized, ostracized, and potentially highly vulnerable communities to lose out on critical online resources and information sharing simply due to a lack of anonymity.

In fact, mandating any sharing of personally identifiable information online is going to alienate communities in the United Kingdom that are most wary of government data collection. Historically, many legal immigrant and United Kingdom-born minorities have been alienated by the United Kingdom government in a myriad of scandals such as Windrush and the Nationality and Borders Act that either used or can use their personal information and immigration status against them. Even if there is no data leak, these minority groups will be afraid to provide more personally identifiable information where mandated by the government.

In 2018, the United Kingdom came under fire for the Windrush scandal, wherein the United Kingdom government made deportation threats to United Kingdom residents whose families had arrived from Caribbean British colonies between the 1940s and 1970s.[105] The United Kingdom government instituted a Windrush compensation scheme for those who lost jobs and housing and were wrongfully deported, but failed to quickly provide those eligible with compensation or payment, further perpetuating fears in these minority groups of the government.[106]

In April 2022, the United Kingdom enacted the Nationality and Borders Act, which could strip people of their United Kingdom citizenship without notice.[107] That same month, the United Kingdom announced the Rwanda migration partnership, wherein asylum seekers (regardless of nationality) would be relocated to Rwanda from the United Kingdom if their protection claims were rejected.[108]

This combination of legislation has created tensions with migrant communities and immigrants in the United Kingdom, leading many legal immigrants and United Kingdom-born minorities to feel unsafe with government mandates. Using government identification as a means of age assurance or age verification is only going to perpetuate these concerns and further alienate communities that fear government surveillance and abuse.

AN OVERBROAD SCOPE OF SERVICES AFFECTED PERPETUATES THESE PRIVACY VIOLATIONS

While the effect of the Online Safety Bill's provisions recommending age assurance measures would be deleterious, the broad scope of the Online Safety Bill only exacerbates the issue.

As previously mentioned, the Online Safety Bill covers more than just social media platforms. User-to-user services are exempt if their only user-created content is emails, SMS messages, MMS messages, aural communication, reviews, or some combination of these types of content, so long as these services have links with the United Kingdom and do not contain pornographic content.[109] Links with the United Kingdom are defined as a "significant number of United Kingdom users" or if the United Kingdom is a "target market" for the online service.[110]

This is critical to understand how the Online Safety Bill perpetuates further privacy concerns. Its scope covers something largely bucketed as protected communications: over-the-top messaging platforms such as Signal, Wire, and WhatsApp that use end-to-end encryption to conceal user-to-user content over the Internet. Protected communications such as these that use end-to-end encryption do so to maintain the privacy and security of both the sender and receiver to make sure that only the participants directly messaging each other can consensually access the content.[111] This mechanism protects the privacy of users of the aforementioned marginalized communities, making them feel safe from threats of persecution, domestic violence, war crimes, and more.

But if the duties of care within the Online Safety Bill overarchingly cover these platforms, then the duties of care must be complied with by these platforms. That means, for all online services (Category 1 and Category 2), preventing users from accessing "priority illegal content" such as terrorism content and CSEA.[112] In risk assessments, it would be impossible for these online services to definitively tell whether such content is being sent through their services. But in Ofcom's risk assessment, if the agency considers there to be a material risk of this content on the platform and Ofcom considers it necessary, Ofcom can further compel the prevention of this content by forcing online services to "use accredited technology to identify … and swiftly take down" terrorism and CSEA content.[113] This would force these private communications platforms to weaken their protections, begin client-side scanning, or create a backdoor foreign adversaries and others could exploit in cyberattacks. But this introduces significant privacy vulnerabilities and the potential outing of those relying on anonymity. And, as Tim Cook famously said, "The reality is if you put a backdoor in, that backdoor's for everybody, for good guys and bad guys."[114]

These duties of care also forget that online services can balance both privacy and child safety, with 29.1 million reports issued by online services to the National Center for Missing and Exploited Children in 2021.[115] Many protected communications such as WhatsApp use techniques that enable user reporting to service providers and balance the security guarantees of the service.[116] WhatsApp even removes 300,000 accounts monthly linked to child predation through metadata analysis, user reporting, and analysis of user profiles.[117] Both service and user reports ensure that enforcement entities can find perpetrators and shield children from harm, and online services can identify and remove harmful content for children online. If forced to give up their security protections to remove this content and awful behavior, online services that use encryption are more likely to leave the United Kingdom altogether. Head of WhatsApp Will Cathcart has gone on record publicly as refusing to kowtow to the mandates within the Online Safety Bill, as WhatsApp has a strong track record of balancing online safety and the public desire for privacy.[118]

HOPE FOR THE ONLINE SAFETY BILL

The Online Safety Bill, while overbroad in scope, does address the credible need for maximized user safety online. However, the legislation is out of balance and needs recalibration in order to create public benefit for both United Kingdom users and international users otherwise extraterritorially impacted.

HOW TO AMEND THE BILL TO BEST BALANCE LEGAL FREE EXPRESSION AND ONLINE SAFETY

The United Kingdom Parliament should amend the Online Safety Bill to clearly define what content is or isn't in scope. Doing so would protect legal free expression and better balance legal free expression with user safety.

Parliament could fix this in one of three ways.

Move Content From Lawful to Unlawful

Amend the bill only to restrict unlawful content and include provisions that move certain types of content from the lawful to unlawful category.

The Online Safety Bill already criminalizes four types of content: newly minting criminal offenses for harmful communications, false communications, threatening communications, and cyberflashing.[119] While harmful, false, and threatening communications fall under the umbrella of broad yet vague definitions within the Online Safety Bill, using the criminalization of cyberflashing as a template can ensure that egregious online harms are taken seriously. In fact, moving types of content from the lawful to unlawful category in a similar way to how the bill criminalizes cyberflashing is likely to clarify what content services must remove and

protect users from without questions of context, nuance, or otherwise.

What makes the criminalization of cyberflashing different from other definitions within the Online Safety Bill is its clear definition, clear application, and clear negative harms.

The Online Safety Bill clearly defines cyberflashing as when a person "intentionally sends or gives a photograph or film of any person's genitals to another person" and intends that the recipient is caused "alarm, distress or humiliation" or sends the content "for the purpose of obtaining sexual gratification and is reckless as to [if the recipient] will be caused alarm, distress or humiliation."[120]

The Online Safety Bill clearly applies the definition by inserting cyberflashing—the dissemination of unsolicited sexual imagery—to the Sexual Offences Act 2003.[121] The Sexual Offences Act 2003 delineates what sexual abuse offenses prosecutors can criminally prosecute individuals for.[122]

Further, this definition ensures that there is clear proof of negative harms. The criminalization of cyberflashing and its inclusion through the Online Safety Bill into the Sexual Offences Act 2003 is in response to clear public outrage over a new and Internet-driven form of sexual abuse. Not criminalizing this would have allowed sexual predators to remain online, even despite their content being potentially moderated. By criminalizing the offense, the United Kingdom is able to mitigate severe cases of Internet harm to adults and children by these predators.

Using a similar test of ensuring clear definition, application, and proof of negative harms will ensure that only content that poses the highest "material risk of significant harm to an appreciable number" of United Kingdom residents is removed and prosecuted to the full extent of the law. Similarly, moving content from the legal but harmful to the unlawful category will force a discussion of proportionate affects to legal free expression that would not be possible if left to the jurisdiction of the politically appointed secretary of state for DCMS or Ofcom.

Prevent Subjective Use of Legal But Harmful

Clearly define what legal but harmful content online services should monitor and remove and prevent subjective amendments in the future.

If Parliament wants to maintain the category of legal but harmful content, it needs to ensure that it's not an overly vague standard that will be left to the discretion twice over of both the secretary of state for DCMS and Ofcom.

The United Kingdom government can achieve consistency and clarity within

the Online Safety Bill by creating a system of periodic Parliament-driven review and amendments to the regulated content. Doing so would ensure that the United Kingdom law does not evolve with political ties or pressures but instead with the rate of Internet innovation and security threats. A systematic review process will also ensure that a fairer balance between legal free expression and online safety is struck with more thought leaders—not just Ofcom and the secretary of state for DCMS—able to discuss what these amendments should look like.

Alternatively, the bill can achieve definitional clarity through a rewrite of content deemed harmful to children by the secretary of state for DCMS or content that falls into the overbroad definition of "significant harm to an appreciable number of" children or adults in the United Kingdom. By rewriting what defines the safety duties for children and safety duties for adults through amendments, Parliament can minimize potential over-moderation by online services trying to follow Ofcom's codes of practice or attempting content moderation risk assessments.

Greater definitional clarity and a structured review and amendment process will enable services to strike a fairer balance between legal free expression and online safety.

Create Broad Immunity Intermediary Liability Protections

Clearly codify intermediary liability protections for online services to proactively moderate content.

Similar to how the United Kingdom originally fell under the European Union's e-Commerce Directive—a law that tackled the basics of transparency requirements and intermediary liability for e-commerce and online services—the United Kingdom could create a broad immunity intermediary liability framework of its own.[123] Broad immunity intermediary liability frameworks protect online services and Internet users from general government-compelled monitoring obligations such as the prescriptive duties of care recommending the use of "proactive technology" within the Online Safety Bill.[124] These frameworks do so by placing the burden of liability on the creators of content, encouraging online services to practice content moderation, and still holding those responsible for illegal content accountable.[125] Codifying sufficient intermediary liability protections will strike a balance that prioritizes both online safety and free expression.

HOW TO AMEND THE BILL TO BEST BALANCE ANONYMITY AND ONLINE SAFETY

With regards to anonymity, a solution that fails to understand the needs of marginalized communities and instead treats age assurance as a one-size-fits-all solution will only uphold the safety of some users while undermining the privacy and safety of others. With no currently available solutions to

age verification that do not reveal or require personally identifiable information, the Online Safety Bill's desire for age assurance is currently incompatible with user privacy. Future legislative measures such as electronic IDs that can be implemented without undermining user privacy or their personally identifiable information are almost final or operational.[126] Until then, users under an Online Safety Bill regime will no longer be able to simply read a website, share content online, or seek information without first proving some facet of their identity. As previously discussed in this paper, this approach has far-reaching consequences.[127]

Therefore, Parliament should amend the Online Safety Bill's approach to age assurance such that services may better balance user safety and user privacy.

Remove Government Prescriptions for Age Assurance

Remove age assurance recommendations that require disclosing personally identifiable information from the proposal and prevent Ofcom from prescribing this technology in the future.

Ensure that United Kingdom adults will not have to disclose personally identifiable information in order to access or share content that previously they could access and share anonymously or semi-anonymously. Studies show that anonymity motivates individuals to speak up and contribute to meaningful democratic discussion.[128] Society can't praise the Arab Spring and other social movements like it that were in part forged through free and anonymous legal expression online, and then turn around and dissolve the very features that made those movements possible. Anonymity is essential for an Internet that functions freely and openly. The perceived end result of safety promised by age assurance simply doesn't justify the means of reduced accountability and violations of personal privacy.

Protect Encryption and Private Communications

Remove protected communications from the scope of the Online Safety Bill.

Excluding messaging services from the list of covered user-to-user services will help ensure United Kingdom users have a right to private communications online. As previously mentioned, the Online Safety Bill already excludes certain communications such as email, SMS messages, and MMS messages from its duties of care placed on user-to-user services.

The inclusion of these exemptions implies legislators have some consideration for the sanctity of private communication. Unfortunately, individuals rely on more than just these three mediums to carry out private messaging. In fact, many users routinely rely on messaging apps that offer end-to-end encryption.[129] For instance, Apple product users benefit from the end-to-end encryption of iMessages sent between two Apple devices

connected to Wi-Fi or cellular data.[130] Previously mentioned apps such as WhatsApp, Wire, Wickr, and Signal are also encrypted messaging platforms that work across devices and have collectively seen a rise in usage.[131] Removing these protected communications from the scope of the bill will ensure the safety of vulnerable individuals who rely on these encrypted mediums to carry out journalism, activism, and other high-risk work. The removal would further protect all other individuals who simply wish to communicate through trusted, safe, and private channels.

Proposed changes to the bill should also explicitly note within the legislative text that nothing in the Online Safety Bill discourages online services from using end-to-end encryption or requires client-side scanning.

CONCLUSION

While the Online Safety Bill currently fails to balance legal free speech, privacy, and online safety, there is hope for a legislative solution that tackles the very real issues Internet users in the United Kingdom and around the world face. Amending the Online Safety Bill should not just be a possibility but a necessity if the United Kingdom is serious about becoming a world leader in Internet safety and online platform regulation.

REFERENCES

1. Big Brother Watch, "#Free Speech Online," accessed July 26, 2022, https://bigbrotherwatch.org.uk/campaigns/freespeechonline/.
2. 47 U.S.C. § 230 (1996). European Parliament and European Council, "Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce')," June 8, 2000, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32000L0031.
3. Charles Hymas, "Loophole in online safety laws 'will let incels spread extremist views,'" *The Telegraph*, April 13, 2022, https://www.telegraph.co.uk/news/2022/04/13/loophole-online-safety-laws-will-let-incels-spread-extremist/.
4. Department for Digital, Culture, Media & Sport and Home Office, "Online Harms White Paper," April 8, 2019, https://www.gov.uk/government/consultations/online-harms-white-paper; Online Safety Bill Draft, Bill CP 405, House of Commons (Sessions 2021–22), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf.
5. Regulatory Policy Committee, "Online Safety Bill: Impact assessment," January 31, 2022, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1061265/Online_Safety_Bill_impact_assessment.pdf.
6. Geoffrey Manne and Ben Sperry, "Warren Bill Highlights the Tradeoffs Inherent in Section 230 Reform," RealClearPolicy, March 25, 2022, https://www.realclearpolicy.com/articles/2022/03/25/warren_bill_highlights_the_tradeoffs_inherent_in_section_230_reform_823570.html.
7. Mikolaj Barczentewicz and Matthew Lesh, "In harm's way: Why online safety regulation needs an Independent Reviewer" (London: Institute for Economic Affairs, February 2022), https://iea.org.uk/publications/in-harms-way-why-online-safety-regulation-needs-an-independent-reviewer/.
8. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 3.
9. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 2.
10. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 190.
11. Ibid.
12. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Schedule 1.
13. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 67(4).
14. Nicolas Wellmann, "OTT-Messaging and Mobile Telecommunication: A Joint Market? – An Empirical Approach" (Düsseldorf Institute for Competition Economics, July 2017), https://www.econstor.eu/bitstream/10419/162779/1/893137103.pdf.

15. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 8.

16. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 52.

17. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 52(7).

18. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Schedule 7.

19. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 52(4).

20. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 9.

21. Markus Trengove, et al., "A Digital Duty of Care: A Critical Review of the Online Safety Bill," April 1, 2022, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4072593.

22. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Sections 10 and 11.

23. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 53.

24. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 55.

25. Ibid.

26. Information Commissioner, "Information Commissioner's opinion: Age Assurance for the Children's Code," Information Commissioner's Office, October 14, 2021, https://ico.org.uk/media/4018659/age-assurance-opinion-202110.pdf.

27. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 54. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 55.

28. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Sections 12 and 13.

29. Ibid.

30. Xuyan Zhu, "Online Safety Bill — illegal and harmful content and safety duties," Taylor Wessing, June 13, 2022, https://www.taylorwessing.com/en/interface/2022/the-online-safety-bill--the-uks-answer-to-addressing-online-harms/online-safety-bill-illegal-and-harmful-content-and-safety-duties.

31. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Sections 12 and 13.

32. Xuyan Zhu, "Online Safety Bill — illegal and harmful content and safety duties," Taylor Wessing, June 13, 2022.

33. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Sections 12-16.

34. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 34.

35. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 82.

36. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 35.

37. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 1.

38. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 37–39.

39. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 45.

40. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 40.

41. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 44.

42. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Part 7.

43. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Schedule 13.

44. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 94.

45. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Sections 8, 10, 12, 23, and 25.

46. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 84.

47. Department for Digital, Culture, Media, and Sport, "Government response to the Joint Bill Committee draft Online Safety Bill," March 17, 2022, https://www.gov.uk/government/publications/joint-committee-report-on-the-draft-online-safety-bill-government-response/government-response-to-the-joint-committee-report-on-the-draft-online-safety-bill.

48. Department for Digital, Culture, Media & Sport and The Rt Hon Nadine Dorries MP, "World-first online safety laws introduced in Parliament," March 17, 2022, https://www.gov.uk/government/news/world-first-online-safety-laws-introduced-in-parliament.

49. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 55.

50. Lexico Dictionaries, s.v. "Material (adj)," accessed August 2, 2022, https://www.lexico.com/en/definition/material; Lexico Dictionaries, s.v. "Significant (adj)," accessed August 2, 2022, https://www.lexico.com/en/definition/significant; Lexico Dictionaries, s.v. "Appreciable (adj)," accessed August 2, 2022, https://www.lexico.com/en/definition/appreciable.

51. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 55.

52. Colm Britchfield, et al., "Government ministers", Institute for Government, accessed July 27, 2022, https://www.instituteforgovernment.org.uk/explainers/government-ministers.

53. Ibid.

54. Charles Hymas, "Loophole in online safety laws 'will let incels spread extremist views'," *The Telegraph,* April 13, 2022.

55. Matthew Lesh and Victoria Hewson, "An Unsafe Bill: How the Online Safety Bill threatens free speech, innovation and privacy" (London: Institute for Economic Affairs, June 2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4172955.

56. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Part 7.

57. Evelyn Douek, "The Administrative State of Content Moderation," Stanford Cyber Policy Center, October 26, 2021, https://www.youtube.com/watch?v=RGdeV4KChWE.

58. Esteban Ortiz-Ospina, "The Rise of Social Media," Our World in Data, September 18, 2019, https://ourworldindata.org/rise-of-social-media.

59. "Britain's Online Safety Bill could change the face of the Internet," *The Economist*, May 25, 2022, https://www.economist.com/britain/2022/05/25/britains-online-safety-bill-could-change-the-face-of-the-Internet; Lisa Hodge, "Scots TikTok mum banned from app for breastfeeding baby slams social media service for 'shaming mothers,'" *Daily Record*, October 2, 2020, https://www.dailyrecord.co.uk/news/scottish-news/scots-tiktok-mum-banned-app-22774757; John Samples, "Why the Government Should Not Regulate Content Moderation of Social Media" (Cato Institute, April 2019), https://www.cato.org/sites/cato.org/files/pubs/pdf/pa_865.pdf.

60. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Sections 9, 11, 13, 24, and 26.

61. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Sections 15 and 16.

62. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Sections 16.

63. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 67(4).

64. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Sections 15.

65. Lexico Dictionaries, s.v. "Citizen Journalism (n)", accessed July 20, 2022, https://www.lexico.com/en/definition/citizen_journalism; "Freedom of speech requirements, journalism, and content of democratic importance," accessed 20 July, 2022, https://publications.parliament.uk/pa/jt5802/jtselect/jtonlinesafety/129/12910.htm#footnote-264.

66. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 50.

67. Online Safety Bill Draft, Bill CP 405, House of Commons (Sessions 2021–22), Explanatory Notes, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985031/Explanatory_Notes_Accessible.pdf; Online Safety Bill, Bill 285, House of Commons (Sessions 2021–22, 22–23), Explanatory Notes,

https://publications.parliament.uk/pa/bills/cbill/58-02/0285/210285en.pdf.

68. Ibid.

69. Ibid.

70. Sam Levin, “Facebook temporarily blocks Black Lives Matter activists after he posts racist email,” *The Guardian*, September 12, 2016, https://www.theguardian.com/technology/2016/sep/12/facebook-blocks-shaun-king-black-lives-matter; “Timeline: Censorship of Pro-life speech by Big Tech,” Susan B. Anthony Pro-Life America, accessed July 28, 2022, https://sbaprolife.org/censorship.

71. Aamna Mohdin and Jessica Murray, “The Mark Duggan case was a catalyst’: the 2011 England riots 10 years on,” *The Guardian*, July 30, 2021, https://www.theguardian.com/uk-news/2021/jul/30/2011-uk-riots-mark-duggan.

72. “Mark Duggan death: IPCC 'may have misled journalists,'” BBC, August 12, 2011, https://www.bbc.co.uk/news/uk-england-london-14510329.

73. Carlie Porterfield, “How Sarah Everard’s Disappearance Sparked A Social Media Movement,” *Forbes*, March 11, 2021, https://www.forbes.com/sites/carlieporterfield/2021/03/11/how-sarah-everards-disappearance-sparked-a-social-media-ovement/?sh=6292df822a1d.

74. Sisters Uncut, “Without the right to protest women have everything to lose – the Sarah Everard vigil proved why,” i News, March 15, 2021, https://inews.co.uk/opinion/women-right-protest-policing-crime-bill-sarah-everard-vigil-914008.

75. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 16.

76. Hasan Obaid, “The Impact of Social Media on Social Movements and Public Opinion Formation,” VISIO International for Rights and Development, April 2020 https://www.researchgate.net/profile/Hasan-Obaid-2/publication/356289366_The_Impact_of_Social_Media_on_Social_movements_and_public_opinion_formation/links/61954c6c3068c54fa5f6d875/The-Impact-of-Social-Media-on-Social-movements-and-public-opinion-formation.pdf; Marcia Mundt, et.al. “Scaling Social Movements Through Social Media: The Case of Black Lives Matter,” University of Massachusetts Boston, November 1, 2018, https://journals.sagepub.com/doi/full/10.1177/2056305118807911.

77. Sam Levin, “Facebook temporarily blocks Black Lives Matter activists after he posts racist email,” *The Guardian*, September 12, 2016, https://www.theguardian.com/technology/2016/sep/12/facebook-blocks-shaun-king-black-lives-matter; “Timeline: Censorship of Pro-life speech by Big Tech,” Susan B. Anthony Pro-Life America, accessed July 28, 2022, https://sbaprolife.org/censorship.

78. Kevin Roose, “What Is QAnon, the Viral Pro-Trump Conspiracy Theory?” *The New York Times*, September 3, 2021, https://www.nytimes.com/article/what-is-qanon.html.

79. Jason Wilson, “Falun Gong-aligned media push fake news about democrats and Chinese communists,” *The Guardian*, April 30, 2022

https://www.theguardian.com/us-news/2021/apr/30/falun-gong-media-epoch-times-democrats-chinese-communists;

80. Travis Andrews, "Gab the social network that has welcomed Qanon and extremists figures, explained," *Washington Post*, January 11, 2021 https://www.washingtonpost.com/technology/2021/01/11/gab-social-network/; "Twitter block 70,000 QAnon accounts after US Capitol riot, AP News," January 12, 2021, https://apnews.com/article/twitter-blocks-70k-qanon-accounts-171a5c9062be1c293169d764d3d0d9c8.

81. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 50.

82. Terrorism Act of 2000, 2000 c.11, https://www.legislation.gov.uk/ukpga/2000/11/contents.

83. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 11.

84. Simone van der Hof, "Age assurance and age appropriate design: what is required," London School of Economics, November 17, 2021, https://blogs.lse.ac.uk/parenting4digitalfuture/2021/11/17/age-assurance/.

85. Introduction to the Age appropriate design code, Information Commissioner's Office, https://ico.org.uk/for-organisations/guide-to-data-protection/ico-codes-of-practice/age-appropriate-design-code/.

86. Information Commissioner, "Information Commissioner's opinion: Age Assurance for the Children's Code," Information Commissioner's Office, October 14, 2021, https://ico.org.uk/media/4018659/age-assurance-opinion-202110.pdf.

87. Ibid.

88. Ibid.

89. Ibid.

90. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Sections 53 and 55.

91. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 68.

92. David McCabe, "Anonymity No More? Age Checks Come to the Web," *New York Times*, October 27, 2021, https://www.nytimes.com/2021/10/27/technology/Internet-age-check-proof.html.

93. Ibid.

94. Pavni Diwanji, "How Do We Know Someone is Old Enough to Use our Apps?" Meta, July 27, 2021, https://about.fb.com/news/2021/07/age-verification/.

95. Rachel Abrams, "Target to Pay $18.5 Million to 47 States in Security Breach Settlement," *New York Times*, March 23, 2017, https://www.nytimes.com/2017/05/23/business/target-security-breach-settlement.html; Josh Fruhlinger, "Equifax data breach FAQ: What happened, who was affected, what was the impact?" CSO, February 12, 2020, https://www.csoonline.com/article/3444488/equifax-data-breach-faq-what-happened-who-was-affected-what-was-the-impact.html; Alvin Chang, "The Facebook and Cambridge Analytica scandal, explained with a

simple diagram," *Vox*, May 2, 2018, https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram.

96. Joseph Berstein, "Ashley Madison's $19 'Full Delete' Option Made the Company Millions," *Buzzfeed News*, August 19, 2015, https://www.buzzfeednews.com/article/josephbernstein/leaked-documents-suggest-ashley-madison-made-millions-promis.

97. Ibid,

98. Zak Doffman, "Ashley Madison Hack Returns to 'Haunt' its Victims: 32 Million Users Now Watch and Wait," *Forbes*, February 1, 2020, https://www.forbes.com/sites/zakdoffman/2020/02/01/ashley-madison-hack-returns-to-haunt-its-victims-32-million-users-now-have-to-watch-and-wait/?sh=57087ec35677.

99. Hannah Shewan Stevens, "Would Removing Social Media Anonymity Protect or Threaten our Rights?" Each Other, February 3, 2022 https://eachother.org.uk/would-removing-social-media-anonymity-protect-or-threaten-our-rights/.

100. Michelle Seigal, "For LGBTQ Youth, Truly Equitable Internet Access Requires End-to-End Encryption," January 28th, 2022, https://www.newamerica.org/oti/blog/for-lgbtq-youth-truly-equitable-Internet-access-requires-end-to-end-encryption/.

101. Ibid.

102. Janus Kopfstein, "Social movement need anonymity, but corporations are taking it away," *Aljazeera*, July 1, 2015, http://america.aljazeera.com/opinions/2015/7/social-movements-need-anonymity-but-corporations-are-taking-it-away.html.

103. "Dissidents under authoritarian rule: Staying anonymous yet trustworthy," University of California, Santa-Barbara, January 15, 2019 https://www.sciencedaily.com/releases/2019/01/190115121106.htm

104. Janus Kopfstein, "Social movements need anonymity, but corporations are taking it away," July 1, 2015.

105. "Windrush generation: Who are they and why are they facing problems?" BBC, November 24, 2021, https://www.bbc.co.uk/news/uk-43782241.

106. Diana Johnson MP, "The government is repeating its failings for Windrush victims," Politics Home, March 3, 2022, https://www.politicshome.com/thehouse/article/government-repeating-windrush-failings.

107. Home Office, "Nationality and Borders Bill: Deprivation of citizenship factsheet," March 2, 2022, https://www.gov.uk/government/publications/nationality-and-borders-bill-deprivation-of-citizenship-factsheet/nationality-and-borders-bill-deprivation-of-citizenship-factsheet.

108. Home Office News Team, "Factsheet: Migration and Economic Development Partnership," April 14, 2022, https://homeofficemedia.blog.gov.uk/2022/04/14/factsheet-migration-and-economic-development-partnership/.

109. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Schedule 1.

110. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 67(4).

111. Wire, “Why Wire?” accessed July 15, 2022, https://wire.com/en/; Signal, “Technical information: specifications and software libraries for developers,” accessed July 15, 2022, https://signal.org/docs/; WhatsApp Messenger, “About end-to-end encryption,” accessed July 15, 2022, https://faq.whatsapp.com/general/security-and-privacy/end-to-end-encryption/?lang=en.

112. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Sections 9 and 104.

113. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Section 104.

114. Herb Lin, “’Back Doors for Good Guys Means Back Doors for Bad Guys’—Unpacking Another Claim,” Lawfare, December 22, 2015, https://www.lawfareblog.com/back-doors-good-guys-means-back-doors-bad-guys-unpacking-another-claim.

115. National Center for Missing and Exploited Children, “2021 CyberTipline Reports by Electronic Service Providers (ESP),” 2021, https://www.missingkids.org/content/dam/missingkids/pdfs/2021-reports-by-esp.pdf.

116. Paul Grubbs, Jiahui Lu, and Thomas Ristenpart, “Message Franking via Committing Authenticated Encryption,” Cryptology ePrint Archive, 2017, https://eprint.iacr.org/2017/664.pdf.

117. Matt Burgess, “Police caught one of the web’s most dangerous paedophiles. Then everything went dark,” *Wired*, December 5, 2021, https://www.wired.co.uk/article/whatsapp-encryption-child-abuse.

118. Shiona McCallum, “WhatsApp: We won't lower security for any government,” BBC, July 30, 2022, https://www.bbc.co.uk/news/technology-62291328.

119. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Sections 151–157.

120. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Sections 157.

121. Ibid.

122. Sexual Offenses Act 2003, 2003 c.42, https://www.legislation.gov.uk/ukpga/2003/42/contents.

123. European Parliament and European Council, "Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce')," 8 June 2000, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32000L0031.

124. Online Safety Bill, Bill 121, House of Commons (Sessions 2021–22, 2022–23), Sections 187.

125. Folkert Wilman, “The EU’s system of knowledge-based liability for hosting providers in respect of illegal user content - between the e-Commerce Directive and the Digital Services Act,” *Journal of Intellectual Property*, Information Technology, and E-Commerce Law, December 3, 2021, https://www.jipitec.eu/issues/jipitec-12-3-2021/5343. Ashley Johnson and Daniel Castro, “Overview of Section 230: What It Is, Why It Was

Created, and What It Has Achieved" (Information Technology and Innovation Foundation, February 2021), https://itif.org/publications/2021/02/22/overview-section-230-what-it-why-it-was-created-and-what-it-has-achieved/.

126. Department for Digital, Culture, Media & Sport and Julia Lopez, MP "New legislation set to make digital identities more trustworthy and secure," March 10, 2022, https://www.gov.uk/government/news/new-legislation-set-to-make-digital-identities-more-trustworthy-and-secure.

127. Heather Burns, "Age Verification in the Online Safety Bill," Open Rights Group, July 27, 2021, https://www.openrightsgroup.org/blog/age-verification-in-the-online-safety-bill/.

128. "Anonymity Online is Important," Digital Rights Watch, April 30, 2021, https://digitalrightswatch.org.au/2021/04/30/explainer-anonymity-online-is-important/.

129. Will Cathcart, "Encryption Has Never Been More Essential— or Threatened," *Wired*, April 5, 2021, https://www.wired.com/story/opinion-encryption-has-never-been-more-essential-or-threatened/.

130. Chad Warner, "Is Apple iMessage End-to-End Encrypted? It Depends," *Medium*, December 3, 2021, https://warnerchad.medium.com/is-apple-imessage-end-to-end-encrypted-it-depends-8bcdcbd8c89.

131. Jack Nicas et al., "Millions Flock to Telegram and Signal as Fears Grow Over Big Tech," *New York Times*, January 13, 2021, https://www.nytimes.com/2021/01/13/technology/telegram-signal-apps-big-tech.html.

ABOUT THE AUTHORS

Kir Nuthi is a senior policy analyst at the Center for Data Innovation focusing on European digital policy. Previously, she worked as a public affairs manager at NetChoice, where she focused on emerging technology issues surrounding content moderation, competition policy, and the sharing economy. Kir holds an MSc in International Public Policy from University College London and a BA with dual focuses in Economics and Political Science from the University of California San Diego.

Mella Tesfazgi is a policy fellow at the Center for Data Innovation. She is a second-year Master's in Public Policy candidate at Duke University's Sanford School of Public Policy and holds a B.A in Economics from Wake Forest University.

ABOUT THE CENTER FOR DATA INNOVATION

The Center for Data Innovation is the leading global think tank studying the intersection of data, technology, and public policy. With staff in Washington, D.C., London, and Brussels, the Center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as open data, artificial intelligence, and the Internet of Things. The Center is part of the nonprofit, nonpartisan Information Technology and Innovation Foundation (ITIF).

contact: info@datainnovation.org

datainnovation.org